REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From - To)*
MARCH 2012	Conference Paper (PREPRINT)	

4. TITLE AND SUBTITLE

GLOTTAL WAVEFORM ANALYSIS OF PHYSICAL TASK STRESS SPEECH (PREPRINT)

5a. CONTRACT NUMBER
FA8750-09-C-0067

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER
35885G

6. AUTHOR(S)

Keith Godin, Taufiq Hasan, and John H. L. Hansen

5d. PROJECT NUMBER
3188

5e. TASK NUMBER
BA

5f. WORK UNIT NUMBER
AE

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Research Associates for Defense Conversion, Inc.
10002 Hillside Terrace
Marcy NY 13403

Subcontractor: University of Texas at Dallas
800 W Campbell Rd
Richardson, TX 75080

8. PERFORMING ORGANIZATION REPORT NUMBER
N/A

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Air Force Research Laboratory/Information Directorate
Rome Research Site/RIGC
525 Brooks Road
Rome NY 13441-4505

10. SPONSOR/MONITOR'S ACRONYM(S)
N/A

11. SPONSORING/MONITORING AGENCY REPORT NUMBER
AFRL-RI-RS-TP-2012-039

12. DISTRIBUTION AVAILABILITY STATEMENT
Approved for public release; distribution unlimited. PA# 88ABW-2012-1817
Cleared Date: 29 March 2012

13. SUPPLEMENTARY NOTES
This paper was accepted for publication in the Proceedings of Interspeech, Portland, Oregon, 9-13 Sept-2012.
This work was funded in whole or in part by Department of the Air Force contract number FA8750-09-C-0067.

14. ABSTRACT
Physical task stress affects the acoustic speech wave in various ways. Motivated by observations that fundamental frequency and open quotient are affected by physical task stress, this study examines the effects of physical task stress on parameters of the estimated glottal volume velocity waveform. It is shown that, in contrast to other types of phonation such as soft, loud, breathy, or pressed voice, physical task stress has little effect on the glottal waveform parameters chosen for analysis. Further, the use of glottal waveform parameters in a stress detection system does not improve the system accuracy, again in contrast to other types of non-neutral speech. These results suggest that a medium level of physical task stress does not greatly perturb vocal fold behavior, and the search for explanations for the spectral perturbations that make stress detection possible must turn to other directions.

15. SUBJECT TERMS
Speech under stress, Tactical SIGINT Technology, Glottal waveform, audio analysis

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			JOHN G. PARKER
U	U	U	UU	5	19b. TELEPHONE NUMBER (*Include area code*) N/A

Glottal Waveform Analysis of Physical Task Stress Speech

Keith W. Godin, Taufiq Hasan, and John H. L. Hansen

Center for Robust Speech Systems
University of Texas at Dallas, Richardson, TX, U.S.A.
godin@ieee.org, taufiq.hasan@utdallas.edu, john.hansen@utdallas.edu

Abstract

Physical task stress affects the acoustic speech wave in various ways. Motivated by observations that fundamental frequency and open quotient are affected by physical task stress, this study examines the effects of physical task stress on a set of glottal features. It is shown that a set of six glottal features can be used for physical stress detection, implying that physical task stress affects vocal fold behavior. It is also shown that the distributions of these six glottal features, across all available speech data, are not affected by physical task stress, leading to the conclusion that covariation in the features due to physical task stress represents different behavioral responses to physical stress. Age and exertion level are explored and rejected as explanatory variables for behavioral types. It is shown, however, that the glottal measurements show changes in distribution when restricted to one recording session of one speaker, suggesting that a given speaker may adopt and retain a particular response to physical task stress for the duration of a task.

Index Terms: glottal waveform, physical task stress, stress detection

1. Introduction

The inference of emotional and physical states of speakers by analysis of the speech signal is of great interest. For example, in the course of the performance of their duties, brush firefighters undergo extreme physical demands. A summary of the overall physical state of their personnel, their task demands, and fatigue levels, through analysis of a variety of physiological signals as well as the speech signal, might one day prove to be invaluable adjunct information to central staff who must decide where to spend limited human firefighting resources. Thus it is of value to catalog the effects of physical task stress on speech and to develop systems that can infer the influence of physical task stress on speakers.

The effects of physical task stress on the speech production system are often subtle. Some of them are known in the literature. Fundamental frequency (F0) increases for most speakers [1] while the variance of F0 does not increase [2]. The glottal open quotient decreases in physical task stress [3]. The center frequencies of the first two formants decrease in certain contexts [3]. There is evidence that high vowels are affected differently than low vowels [4], and that nasals are affected more than fricatives and plosives [4].

Recent advances in methods for the estimation of the glottal volume velocity waveform have added to an already large catalog of methods for analysis of the acoustic consequences of changes in vocal fold behavior [5]. Various types of non-modal phonation are known to have a distinct effect on glottal waveform parameters, such as loud and soft speech [5], and breathy and pressed phonation types [6] [7]. Hypothetically, physical task stress could result in at times breathy and at other times pressed phonation. Given evidence that the glottal open quotient decreases due to physical task stress [3], the research question posed in this study is whether the effects of physical task stress manifest as extensive changes in glottal volume velocity waveform parameters, as observed for other speech types.

A stress detection system is first used to establish the dependence of six glottal waveform parameters on physical task stress. These six parameters are then measured and compared in physical task stress and neutral speech.

2. Corpus

The physical task stress data used in this study is drawn from the first and second physical task stress collects of the UT-Scope corpus [8] [3]. The first and second stages of the collect (in 2007 and 2010, respectively) were collected in the same sound booth, using the same recording equipment and microphone, on the same exercise machine, and both include a 35-sentence prompted speech segment in both physical stress and neutral tasks. This study uses the 35-sentence prompted segment, drawn from 70 female native speakers of American English. Some speakers spoke multiple sessions, and some utterances were repeated due to pronunciation mistakes, resulting in a total of 2815 neutral utterances and 2717 physical stress utterances. The collection paradigm subjects each participant to the same physical task, resulting in a different level of exertion across speakers, depending on their physical fitness. For most speakers, the task resulted in a medium level of exertion, one in which they could comfortably speak [4].

3. Stress detection

The purpose of the following experiment is to determine whether physical task stress affects the glottal waveshape. Systems for physical task stress detection [9] use MFCCs and TEO-CB-AutoEnv [10] as acoustic features, and a pair of GMMs as classifier. This experiment tests glottal waveshape features. The six features used are discussed in Section 4. They are mostly dependent on the glottal waveshape and thus on vocal fold behavior.

The detection task is to determine whether or not a given 2-3 second sentence was produced in the context of physical task stress. The stress detection system uses the log likelihood ratio between a pair of 512 mixture GMMs as classification statistic. Several speakers recorded two sessions and all recorded data

This project was funded by AFRL under contract FA8750-12-1-0188 (Approved for public release; distribution unlimited: 88ABW-2012-1817), and partially by the University of Texas at Dallas from the Distinguished University Chair in Telecommunications Engineering held by J.H.L. Hansen.

Table 1: Detection accuracy as a function of acoustic feature. Chance level is 50% accuracy.

System	Accuracy (%)
MFCCs	74.4
6 glottal features	69.0
Feature fusion	74.0
System fusion	74.1

was included in the evaluation task. Some speakers repeated utterances because of pronunciation mistakes, resulting in different numbers of physical task stress and neutral utterances. The training data consists of 1,695 neutral utterances, and 1,632 physical task stress utterances, from 45 speakers. The evaluation data consists of 1,085 physical task stress utterances and 1,120 neutral utterances, from 25 speakers not included in the training data. The baseline feature is 13 dimensional MFCCs with deltas and double-deltas, 25 ms frames and 10 ms frame hop.

The stress detection task here is not directly comparable to [9]. In the present study, the detection utterances have all been carefully hand-clipped to ensure that no recorded breaths are contained. This ensures that the detection task is concerned with the speech signal, rather than breaths, because the glottal parameter analysis discussed in Section 4 is concerned solely with the speech signal.

Table 1 shows the detection accuracy results for each feature set. The system using the 6 glottal features achieved a 69% detection accuracy. This suggests that physical task stress affects the glottal waveshape in detectable ways. This is not surprising, given previous analysis results [3], and given that a cognitive stress detection system using similar glottal waveshape parameters resulted in an above-chance detection accuracy [11]. However, what is surprising is that neither feature level fusion nor system level fusion with MFCCs results in an improvement over the MFCC detection accuracy. This is inconsistent with results for cognitive stress, in which a 6% absolute improvement in detection accuracy for fusion was observed [11]. Section 4 explores the effects of physical task stress on these six glottal waveshape parameters in more detail.

4. Glottal and acoustic speech waveform parameters

Section 3 showed that a set of six glottal features are dependent on physical task stress. These features are discussed in this section and are the subject of an analysis. For other types of non-neutral voicing including pressed, breathy, soft, and loud, an average across all frames of all speakers has been sufficient in the literature to demonstrate that glottal parameters are dependent on the speech type under study [7] [5]. This section demonstrates that this is not the case for physical task stress, and proceeds to explore age, exertion level, and speaker identity, in the search for a grouping of responses that might demonstrate a dependence of the parameters on physical task stress.

Three of the parameters are measured from an estimated glottal volume velocity waveform; the remaining three are measured directly from the acoustic wave. The GLOAT toolkit [5] is used to estimate the glottal volume velocity waveform derivative; an integrating filter is used to estimate the volume velocity waveform. The GLOAT toolkit is used to select only voiced frames for speech analysis, and to estimate the fundamental frequency. Analysis frames of 25 ms are used, with a fixed 10 ms frame skip.

4.1. Harmonics to Noise Ratio

The Harmonics to Noise Ratio (HNR) [12] was developed as a measure of hoarseness of a speaker. The HNR as originally defined is computed by averaging in the time domain across a window of nearby pitch periods to obtain the average periodic component, and then subtracting this average from one pitch period to obtain the noise component, and taking the ratio of the energy of the periodic component to the energy of the noise component. A cepstral method is used here to estimate the HNR [13].

4.2. F1F3syn

F1F3syn is designed to detect aspiration noise at the glottis [15]. It is computed from the cross-correlation of the low-pass filtered Hilbert envelopes of two wide frequency bands (100Hz-1.5kHz, 1.8-4kHz). If there is more aspiration noise than in typical neutral speech, the harmonic structure in the two bands will be less correlated and the F1F3syn will be lower.

4.3. Normalized Amplitude Quotient

The Normalized Amplitude Quotient (NAQ) is the ratio of the maximum amplitude of the glottal flow to the minimum of the glottal flow derivative, normalized by the fundamental period and the sampling frequency [7]. NAQ is sensitive to variations caused by breathy and pressed phonation [7] and to soft and loud speech [5]. NAQ increases for breathy phonation and decreases for pressed phonation, relative to neutral speech.

4.4. Harmonic Richness Factor

The Harmonic Richness Factor (HRF) is the ratio of the sum of the amplitudes at the harmonics in the glottal waveform to the amplitude of the component at the fundamental frequency [16]. In [16], the HRF of modal voicing was higher than that for breathy voicing by 6.8 db. In [5], there were clear shifts in the distribution of HRF between loud, modal, and soft voicing.

4.5. H1H2 Ratio

The H1H2 ratio is the ratio between the amplitude at the first harmonic to the amplitude at the second harmonic [17]. In that study, an increase in the H1H2 ratio was correlated with perceived breathiness.

4.6. Spectral Slope

The spectral slope is well known as the slope of the least-squares regression line fitted to the log-magnitude spectrum bins for one speech frame, and is measured in decibels. Angry, loud, and Lombard effect speech significantly affect the spectral slope [14]. The spectral slope for voiced frames is typically negative. If the glottal waveshape is less smooth than typical neutral speech, it will have a stronger harmonic structure, and a spectral slope closer to 0.

4.7. Global analysis

All frames of all speakers were used to create a pair of histograms for each of the six parameters, shown in Figure 1. These plots stand in sharp contrast to the results from [5], where glottal parameters from loud and soft speech were found to vary significantly from neutral when a histogram was plotted across

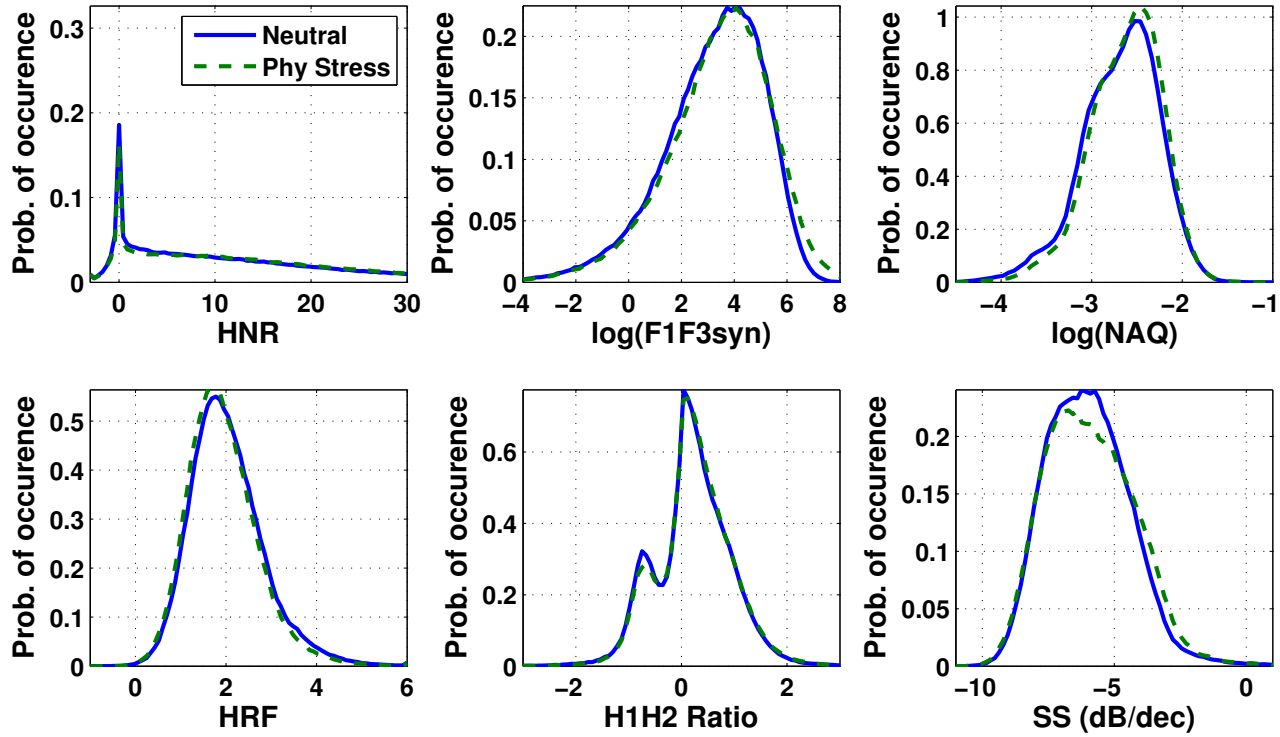

Figure 1: Parameters in physical task stress and neutral, across all voiced frames for all speakers.

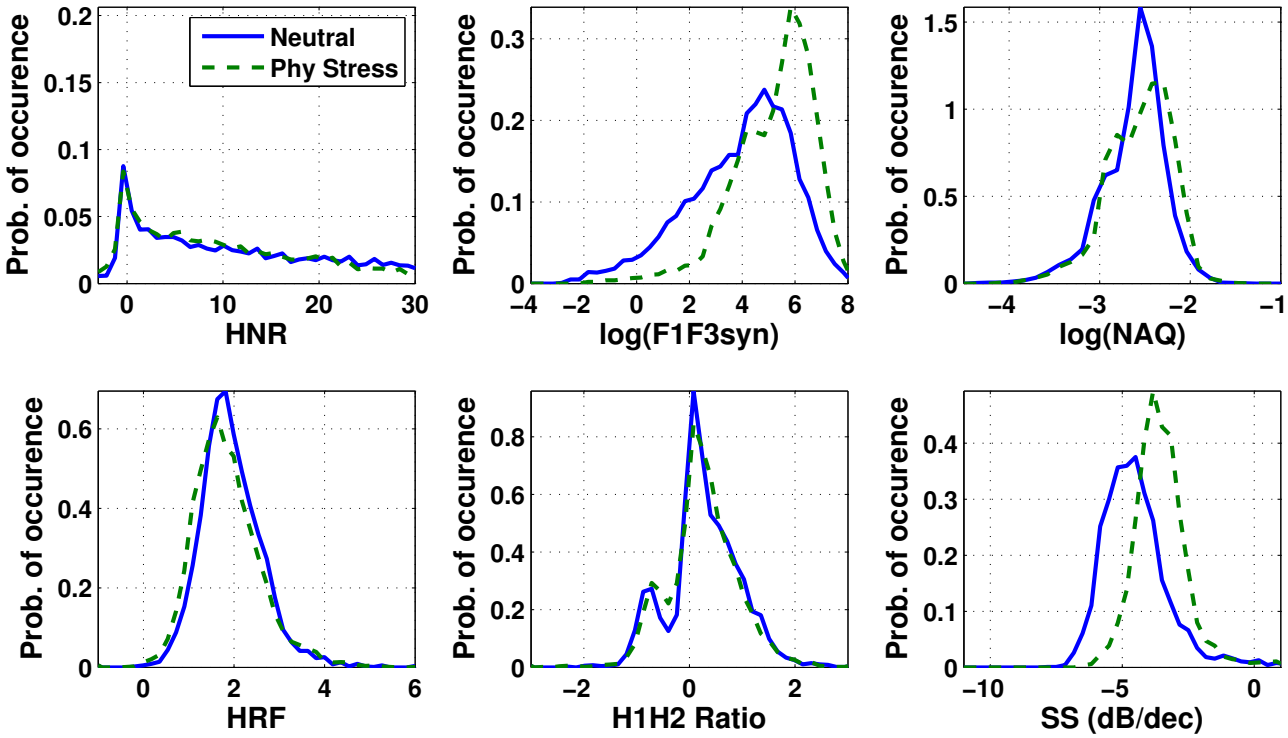

Figure 2: Parameters in physical task stress and neutral, across all voiced frames for speaker VENSF003.

all speakers, as here. These plots do not suggest that physical task stress does not affect the glottal waveshape, because Section 3 shows that these six features can be used to detect physical task stress. Instead, it appears that there are many different responses to physical task stress represented by covariation in the 6 parameters, and while they can be clustered and detected with a GMM, these separate responses average out across all recordings.

4.8. Grouping by age

To explore possible explanations for this clustering, the speakers are clustered according to age. Three groups are formed: speakers aged 18-19, speakers aged 21-30, and speakers aged 40+. In the interest of space, the plots are not shown, rather they are equivalent to the plots of Figure 1. This suggests that speaker age is not a direct explanatory factor for different vocal fold behavioral responses to physical task stress.

4.9. Grouping by exertion level

[4] showed a significant difference in a spectral measure between one of four exertion levels. Exertion level of a speaker depends on the heart rate in neutral and physical task stress, and the speaker's age. The speakers were grouped by the exertion levels from [4] and the histograms examined. They are not shown in the interest of space; they are equivalent to those of Figure 1, showing overlapping distributions. As with age, exertion level does not appear to be a direct explanatory factor for differences in vocal fold behavioral responses to physical task stress.

4.10. Individual speaker response

Finally, Figure 2 shows the six measurements averaged across all of the frames of a single speaker. The plots for F1F3syn, NAQ, and SS show apparent differences in distribution. This suggests that there are consistent vocal fold behavioral responses to physical stress at least within a single session of a single speaker's speech. It may be that speakers adopt different responses given different exertion levels or in different sessions. However, this plot does suggest some of the covariation in parameters, which may be associated with a particular behavioral response to physical stress, that make detection of physical task stress possible.

5. Conclusions

This study has shown that physical task stress detection is possible with a set of 6 glottal features. In conjunction with [1], [2], and [3], this leads to the conclusion that physical task stress affects vocal fold behavior. An analysis of the six parameters showed that across all available speech data, the distributions of the six parameters did not depend on physical task stress. Instead, given the classification result, it must be assumed that covariation in the parameters occurs in response to physical task stress. This suggests that there are behaviors that cluster in these six parameters and thus are amenable to modeling by a GMM. An analysis of age and exertion level as explanatory factors did not suggest that they directly result in clustering of vocal fold behaviors. Instead, it was shown that within a single session of a single speaker, vocal fold behavioral responses to physical task stress were consistent enough to shift the distributions of three parameters.

This consistency of response within a given session suggests that if the physical task is constant, the response to stress won't necessarily vary from utterance to utterance. These results suggest future work to determine whether there are just a few overall responses to physical task stress that can be grouped together through some means of clustering the glottal parameters in this study. If so, perhaps a set of exogenous variables can be found to explain the grouping, such as speaker age, fitness level, exertion level, or fatigue level.

6. References

[1] B. Johannes, P. Wittels, R. Enne, G. Eisinger, C. A. Castro, J. L. Thomas, A. B. Adler, and R. Gerzer, "Non-linear function model of voice pitch dependency on physical and mental load," *Eur. J. Appl. Physiology*, vol. 101, pp. 267–276, 2007.

[2] K. W. Godin and J. H. L. Hansen, "Analysis and perception of speech under physical task stress," in *INTERSPEECH 2008*, Brisbane, Australia, Sep. 2008, pp. 1674–1677.

[3] ——, "Vowel context and speaker interactions influencing glottal open quotient and formant frequency shifts in physical task stress," in *Interspeech 2011*, 2011, pp. 2945–2948.

[4] ——, "Analysis of the effects of physical task stress on the speech signal," *J. Acoust. Soc. Am.*, vol. 130, pp. 3992–3998, 2011.

[5] T. Drugman and A. Alwan, "Joint robust voicing detection and pitch estimation based on residual harmonics," in *Interspeech 2011*, 2011, pp. 1973–1976.

[6] P. Alku and E. Vilkman, "A comparison of glottal voice source quantification parameters in breathy, normal and pressed phonation of female and male speakers," *Folia phoniatrica et logopaedica*, vol. 48, pp. 250–254, 1994.

[7] P. Alku, T. Backstrom, and E. Vilkman, "Normalized amplitude quotient for parametrization of the glottal flow," *J. of the Acoustical Soc. of Am.*, vol. 112, pp. 701–710, 2002.

[8] A. Ikeno, V. Varadarajan, S. Patil, and J. H. L. Hansen, "UT-Scope: Speech under lombard effect and cognitive stress," in *IEEE Aerospace Conf. 2007*, Big Sky, Montana, 2007, pp. 1–7.

[9] S. A. Patil and J. H. L. Hansen, "Detection of speech under physical stress: Model development, sensor selection, and feature fusion," in *INTERSPEECH 2008*, 2008.

[10] G. Zhou, J. H. L. Hansen, and J. F. Kaiser, "Nonlinear feature based classification of speech under stress," *Speech and Audio Proc., IEEE Trans. on*, vol. 9, no. 3, pp. 201–216, March 2001.

[11] T. F. Yap, J. Epps, E. H. C. Choi, and E. Ambikairajah, "Glottal features for speech-based cognitive load classification," in *ICASSP 2010*, 2010, pp. 5234 – 5237.

[12] E. Yumoto, W. J. Gould, and T. Baer, "Harmonics-to-noise ratio as an index of the degree of hoarseness," *J. of the Acoustical Soc. of America*, vol. 71, no. 6, pp. 1544–1550, 1982.

[13] G. de Krom, "A cepstrum-based technique for determining a harmonics-to-noise ratio in speech signals," *J. of Speech and Hearing Res.*, vol. 36, pp. 254–266, 1993.

[14] J. H. L. Hansen, "Analysis and compensation of stressed and noisy speech with application to robust automatic recognition," Ph.D. dissertation, Georgia Institute of Technology, Atlanta, July 1988.

[15] C. T. Ishi, "A new acoustic measure for aspiration noise detection," in *Interspeech*, 2004.

[16] D. G. Childers and C. K. Lee, "Vocal quality factors: Analysis, synthesis, and perception," *J. of the Acoustical Soc. of Am.*, vol. 90, no. 5, pp. 2394–2410, Nov. 1991.

[17] D. H. Klatt and L. C. Klatt, "Analysis, synthesis, and perception of voice quality variations among female and male talkers," *J. of the Acoustical Soc. of Am.*, vol. 87, pp. 820–857, 1990.